

Marginalizable conditional model for clustered ordinal data

Rui Zhang, Kwun Chuen Gary Chan

Department of Biostatistics, University of Washington, Seattle, WA, USA

zhangrui@uw.edu, kcgchan@uw.edu

April 1, 2014

ABSTRACT

We introduce a flexible parametric mixed effects model for correlated binary data, with parameters that can be directly interpreted as marginal odds ratios. This leads to a robust estimation equation with an optimal weighting matrix being the inverse of a genuine model-based covariance matrix. Flexible correlation structures can be imposed by correlated random effects, and correlation parameters can be estimated by solving a composite likelihood score function. Marginal parameters are consistently estimated even when the conditional parametric model is misspecified, and the robust estimation procedure has low estimation efficiency loss compared to the maximum likelihood estimation under a correct model specification. Simulations, analyses of the Madras longitudinal schizophrenia study and British social attributes panel survey were carried out to demonstrate our method.

Keywords: alternating logistic regression; complementary log-log link; marginal model; multivariate exponential distribution; mixed effects model.

1 Introduction

Correlation exists naturally when observations are grouped into clusters. For instance, observations are collected from the same subjects at different time points in longitudinal studies. For observations within a cluster, data are typically correlated even after adjusting for observed covariates. We need to address such correlations in a valid statistical analysis. One can often evaluate two distinct covariate effects from clustered data: the *marginal* covariate effect as a population-averaged effect from the study population and the *conditional* covariate effect that quantifies the effect conditional on some unobservable random effects, e.g. cluster-specific effects. Distinct models and methods have been proposed for estimating the marginal and conditional covariate effects. However, marginal and conditional models are typically incompatible for non-linear models such as a logistic regression model. In this paper we consider an unified marginal and conditional model for correlated binary data Y and a vector of covariates X , with (X_i, Y_i) being a vector of observations from the i^{th} cluster and (X_{ij}, Y_{ij}) denoting the j^{th} component/observation, $j = 1, \dots, n_i$. Generalization to three-level clustered data will be discussed in Section 4.

Marginal models were introduced to estimate the marginal covariate effects, which are directly interpretable and are preferred to answer public health questions, according to Neuhaus et al. (1991) and Heagerty (1999). These models are often semi-parametric, which only assume the first and perhaps the second moments of outcomes conditioning on covariates. Under a marginal mean model $E(Y_{ij} | X_{ij}) = g(X_{ij}^T \beta)$ where g is a known inverse link function, the parameter β represents a transformation of the population-average change in expected response per unit change in a given predictor, controlling for the other covariates. For β inference, Liang and Zeger (1986) proposed the Generalized Estimating Equation (GEE). For a dataset containing m independent clusters, the estimate is obtained by solving

$$\sum_{i=1}^m D_i^T V_i^{-1} S_i = 0 ,$$

where $S_i = Y_i - g(X_i^T \beta)$, $D_i = \partial g(X_i^T \beta) / \partial \beta$ and V_i is a "working" covariance matrix given by $V_i = A_i^{1/2} R(\alpha) A_i^{1/2} / \phi$, A_i is a diagonal matrix with elements proportional to $\text{var}(Y_i) = h(X_i^T \beta) / \phi$ and $R(\alpha)$ is a cluster-common working correlation matrix parametrized by α . Nuisance parameters α and ϕ are typically estimated by the method of moments. McCullagh and Nelder (1989) pointed out that the optimal estimation efficiency will be achieved when V_i is the true covariance matrix of Y_i . Note that the working correlation matrix $R(\alpha)$ may not correspond to a genuine correlation matrix from any plausible joint distribution of binary outcomes, as discussed by Chaganty and Joe (2004), who argued that $R(\alpha)$ should be viewed as a weighting matrix, and α should be fixed instead of being estimated.

On the other hand, mixed effect models are commonly used for modeling conditional covariate effects. In general, some unobservable random effects are introduced to model latent cluster effects that cannot be explained by observed covariates and thus together with covariates, they fully characterize correlations between observations; i.e., conditioning on random effects and covariates, observations are assumed to be independent. These models gained popularity because complex correlation structures can be modeled naturally by Gaussian random effects, cluster-specific predictions can be made and likelihood inference is directly applicable. Let b_{ij} denote an unobserved random effect with a conditional density $f(b_{ij} | X_{ij})$. A conditional parametric model specifying the distribution of outcome given observed covariates X_{ij} and random effects b_{ij} is typically assumed. The observed likelihood can be constructed as a marginal density by integrating this conditional outcome density over the random effect distribution:

$$\text{pr}(Y_{ij} | X_{ij}) = \int \text{pr}(Y_{ij} | X_{ij}, b_{ij}) f(b_{ij} | X_{ij}) db_{ij} . \quad (1)$$

However, in general there does not exist a closed-form expression for (1) except for the Gaussian linear mixed model and a few other special cases, causing two problems: 1) observed data likelihood inference requires heavy computation; 2) we cannot directly estimate marginal covariate effects because the lack of a closed-form expression.

Regarding the first problem, numerical integration/approximation techniques have been developed to maximize the observed data likelihood or its approximations, such as the penalized quasi-likelihood inference by Breslow and Clayton (1993), Laplace approximations by Shun and McCullagh (1995), Gauss-Hermite quadrature and Monte Carlo importance sampling algorithms by O'Brien and Dunson (2004). Related methods have been described in details and compared by Pinheiro and Bates (1995).

Several authors have offered solutions to the second problem from different perspectives. From (1), we know that a marginal model and a random effect distribution will jointly determine a conditional model. Likewise, marginal and conditional models will jointly determine the random effect distribution. Heagerty (1999) and Heagerty and Zeger (2000) first jointly modeled the marginal mean model and the random effect distribution and then solved for the conditional mean model, giving the marginalized multilevel models. While this method is conceptually appealing, the implementation is not straightforward since a deconvolution problem is involved, leading to a certain difficulty to the model formulation and interpretation. The bridge distribution proposed by Wang and Louis (2004) started from a fixed pair of marginal and conditional mean models; the authors solved for the random effect distribution and named it the bridge distribution. However, the bridge distribution may not correspond to any known parametric distribution and a lack physical interpretation is also a concern. One may model the joint distribution of a random vector from marginal distributions using a copula, and Song et al. (2009) applied this approach for marginal reference.

Our model formulation for binary data starts from a different perspective. A conditional mean model and a family of correlated random effects are specified to complete the parametric specification of the joint distribution, while directly leading to a marginal logistic regression model. Our formulation is partly motivated from frailty models in survival analysis. The model formulation will be discussed in Section 2. Robust inference is developed in Section 3, extending the generalized estimating equation in Liang and Zeger (1986) and the alternating logistic regression proposed by Carey et al. (1993). The marginal odds ratio parameters can be consistently estimated even when the working conditional mean model or the random effect distribution is misspecified. Asymptotic properties of the estimators are presented in Subsection 3.4. In Section 4, we discuss extensions to three-level clustered data. We show the three-level correlation structure can be naturally incorporated into our model and thus marginal inference can be easily extended into this case. Numerical simulations will be presented in Subsection 5.1, which demonstrate the proposed estimator has a small bias, is robust against model mis-specification and has a negligible efficiency loss compared to maximum likelihood inference. Analyses of the Madras longitudinal schizophrenia study and the British social attributes panel survey will be presented in Subsections 5.2 and 5.3. Concluding remarks and discussions will be given in Section 6. Technical conditions and a proof of the main theorem will be provided in the appendix.

2 A marginalizable conditional model for correlated binary data

2.1 A motivation from frailty models

Our marginalizable mixed effect model is motivated from a close examination of Cox-type frailty models from the survival analysis literature. In these models, given values of frailty a_{ij} and covariate x_{ij} , the conditional hazard rate at time t of the j^{th} observation from the i^{th} cluster is formulated by $a_{ij}\lambda_0(t)\exp(x_{ij}^T\beta)$, where $\lambda_0(\cdot)$ is an unspecified baseline hazard rate function. Its conditional survival probability is

$$S(t \mid X_{ij}, a_{ij}) = \exp\left(-a_{ij}\Lambda_0(t)e^{X_{ij}^T\beta}\right), \quad \text{where } \Lambda_0(t) := \int_0^t \lambda_0(s)ds. \quad (2)$$

The frailties a_{ij} are equivalent to exponentiated random intercepts. For model identifiability in the presence of an unknown baseline hazard rate $\lambda_0(\cdot)$, no intercept term is included into the frailty models and one assumes $E(a_{ij}) = 1$.

It is common to assume the frailty follows a Gamma distribution with mean one and unknown variance $1/\gamma$ to be estimated, with the density

$$f_\gamma(a) = \frac{\gamma^\gamma}{\Gamma(\gamma)} a^{\gamma-1} e^{-\gamma a},$$

see Clayton (1978), Oakes (1982), Hougaard (1984), Vaida and Xu (2000) and Klein (1992) for relevant discussions. Integrating over a_{ij} gives the marginal survival probability

$$S(t | X_{ij}) = \int_0^\infty \exp\left(-a_{ij}\Lambda_0(t)e^{X_{ij}^T\beta}\right) \frac{\gamma^\gamma}{\Gamma(\gamma)} a_{ij}^{\gamma-1} e^{-\gamma a_{ij}} da_{ij} = \left(\frac{1}{1 + \Lambda_0(t)e^{X_{ij}^T\beta - \log\gamma}}\right)^\gamma.$$

Setting $\gamma = 1$, frailties become marginally exponential distributed and the above marginal survival probability simplifies into

$$S(t | X_{ij}) = \frac{1}{1 + \Lambda_0(t)e^{X_{ij}^T\beta}}.$$

Thus at $\gamma = 1$, β can be marginally interpreted as the log failure odds ratio.

2.2 A model for correlated binary data

In the absence of censoring and suppose we are interested in modeling the survival probability at a certain time point t^* , correlated survival outcomes are equivalent to correlated binary outcomes where the binary outcome is $Y_{ij} = I(T_{ij} > t^*)$, where T_{ij} a survival outcome. We assume the conditional probability of the binary outcome follows

$$\text{pr}(Y_{ij} = 1 | X_{ij}, a_{ij}) = \exp\left(-a_{ij}e^{-X_{ij}^T\beta}\right), \quad (3)$$

where a_{ij} 's are marginally standard exponential distributed. In this formulation an intercept is included into the linear predictor, corresponding to $\log\{\Lambda_0(t^*)\}$ from (2).

Similarly to the survival model, the marginal survival probability becomes

$$\text{pr}(Y_{ij} = 1 | X_{ij}) = \frac{1}{1 + e^{-X_{ij}^T\beta}} = \frac{e^{X_{ij}^T\beta}}{1 + e^{X_{ij}^T\beta}}. \quad (4)$$

Therefore marginally, outcomes follow a logistic regression model with the same β coefficients as in the working conditional model (3). We describe the conditional model as a working model, because in Section 3 we will propose a robust estimator for β under the marginal model (4), which is consistent even when the working conditional model (3) is misspecified.

2.3 Random effect variance

Now suppose frailties are exponentially distributed with a variance γ^{-2} , a similar marginalization as in Subsection 2.2 can be obtained, where

$$\text{pr}(Y_{ij} = 1 \mid X_{ij}) = \frac{1}{\gamma} \int_0^\infty \exp\left(-a_{ij}e^{-X_{ij}^T\beta} - \frac{a_{ij}}{\gamma}\right) da_{ij} = \frac{1}{\gamma} \frac{1}{e^{-X_{ij}^T\beta} + \frac{1}{\gamma}} = \frac{1}{e^{-X_{ij}^T\beta + \log\gamma} + 1}.$$

We can see $\log(\gamma)$ merges with the intercept in the marginal probability, implying γ is not identifiable marginally. Moreover, γ is not identifiable in the joint likelihood. For example,

$$\begin{aligned} & \text{pr}(Y_1 = 0, Y_2 = 1, \dots, Y_n = 1) \\ &= \text{pr}(Y_2 = 1, \dots, Y_n = 1) - \text{pr}(Y_1 = 1, \dots, Y_n = 1) \\ &= |I + C_{-1}\text{diag}(\gamma e^{-X_2^T\beta}, \dots, \gamma e^{-X_n^T\beta})|^{-1} - |I + C\text{diag}(\gamma e^{-X_1^T\beta}, \dots, \gamma e^{-X_n^T\beta})|^{-1}, \end{aligned}$$

where C_{-1} is the element-wise square root of the correlation matrix between (a_2, \dots, a_n) and C is the element-wise square root of the correlation matrix between (a_1, \dots, a_n) . Therefore, $\log(\gamma)$ merges with the intercept in joint probabilities as well, and the variance of the random effect cannot be separately estimated from the intercept.

In view of this identifiability problem, we will standardize the random effect distribution having a unit variance.

We note that in conventional linear and logistic mixed models, the within-cluster correlation is controlled by the variance of some shared random effects. While the variance of the random components is standardized in our model, flexible within-cluster correlation can still be modeled by correlated random effects as discussed below, as opposed to using a shared random effect that is often assumed in conventional models.

2.4 Random effect correlation

We allow the frailties to be correlated within clusters and follow a multivariate exponential distribution, instead of assuming frailties are identical within each cluster. To facilitate the modeling of correlations, a class of multivariate exponential distributions can be constructed from multivariate normal distributions as shown in Krishnamoorthy and Parthasarathy (1951) and Henderson and Shimakura (2003). Set W_1 and W_2 to be two independent p -variate, zero-mean and unit-variance Gaussian distributed random vectors; i.e. $W_j = (W_{j1}, \dots, W_{jp})$, $j = 1, 2$. Denote their $p \times p$ correlation matrix by C . Let $Z_k = (W_{1k}^2 + W_{2k}^2)/2$, $k = 1, \dots, p$. For each k , $2Z_k$ is marginally $\chi^2(2)$ distributed; therefore Z_k follows a standard exponential distribution. Moreover, the correlation matrix R of the random vector (Z_1, \dots, Z_p) is an element-wise square of C , see Henderson and Shimakura (2003) for related discussions.

The above connection between multivariate exponential and Gaussian distributions allows one to model flexible correlation patterns similar to the Gaussian mixed effect models. In the following, we will parametrize the correlation matrix for a multivariate exponential random vector by a possibly vector-valued parameter ρ and we will discuss models for three-level clustered data

in Section 4.

2.5 Generalization to covariate-dependent distributed frailties

Although we originally considered the frailty distribution to be independent of covariates, the proposed method for marginal inference is unaffected when covariates are covariate-dependent. Suppose given a covariate X_{ij} , a_{ij} is exponential distributed with mean $e^{X_{ij}^T \gamma}$, and a frailty vector a_i given X_i has a correlation matrix R . Consider the rescaled frailty vector $\tilde{a}_i := (e^{-X_{i1}^T \gamma} a_{i1}, \dots, e^{-X_{in_i}^T \gamma} a_{in_i})$, which is multivariate exponential with mean one and has the same correlation matrix R , the conditional probability of the binary outcome follows from (3), i.e.

$$\text{pr}(Y_{ij} = 1 \mid X_{ij}, a_{ij}) = \exp\left(-a_{ij}e^{-X_{ij}^T \beta}\right) = \exp\left(-\tilde{a}_{ij}e^{-X_{ij}^T \tilde{\beta}}\right), \quad \text{where } \tilde{\beta} = \beta - \gamma.$$

And the marginal probability (4) becomes

$$\text{pr}(Y_{ij} = 1 \mid X_{ij}) = \frac{e^{X_{ij}^T \tilde{\beta}}}{1 + e^{X_{ij}^T \tilde{\beta}}}.$$

For marginal inference, the parameter of interest is $\tilde{\beta}$ and can be estimated as if the frailties were covariate independent.

3 Estimation

3.1 Estimating equation for β with an optimal weighting matrix.

Since the proposed model is parametric, it is natural to consider maximum likelihood estimation (MLE) for model inference, as discussed by Conaway (1990) and Coull et al. (2006). However, MLE has two major drawbacks. First, obtaining consistent MLE requires a correct specification of the conditional model and the random effect distribution, even when the marginal parameters are of main interest. Besides, the likelihood function involves up to $2^n - 1$ terms for each cluster, where n is the cluster size. It may be practically infeasible to compute MLE even for a moderate cluster size, since the computation burden grows exponentially with cluster size.

We propose a robust estimation procedure for the marginal covariate effects β , by replacing the working correlation matrix $R(\alpha)$ in Liang and Zeger's GEE with a real correlation matrix, derived from the conditional model in (3).

Denote the whole set of parameters by $\theta := (\beta, \rho)$. Let g be the inverse of the logit link function:

$$g(x_{ij}^T \beta) := \text{pr}(Y_{ij} = 1 \mid X_{ij} = x_{ij}) = \exp(x_{ij}^T \beta) / (1 + \exp(x_{ij}^T \beta)).$$

For β inference, we solve for

$$\frac{1}{m} \sum_{i=1}^m D(X_i; \beta)^T V^{-1}(X_i; \theta) S(X_i, Y_i; \beta) = 0, \quad (5)$$

where $D(X_i; \beta) = \partial g(X_i^T \beta) / \partial \beta$, $S(X_i, Y_i; \beta) = Y_i - g(X_i^T \beta)$ and $V(X_i; \theta)$ is the $n_i \times n_i$ covariance matrix of the outcome

Y_i . To be more specific, the j^{th} diagonal entry of $V(X_i; \theta)$ is given by

$$V_{jj}(X_i; \theta) = \frac{e^{x_{ij}^T \beta}}{(1 + e^{x_{ij}^T \beta})^2}.$$

Its j^{th} row and k^{th} column entry is

$$V_{jk}(X_i; \theta) = \left[\frac{1}{(1 - \rho_{jk})e^{-(x_{ij} + x_{ik})^T \beta} + e^{-x_{ij}^T \beta} + e^{-x_{ik}^T \beta} + 1} - \frac{1}{1 + e^{-x_{ij}^T \beta}} \frac{1}{1 + e^{-x_{ik}^T \beta}} \right], \quad j \neq k,$$

where ρ_{jk} is the correlation between a_{ij} and a_{ik} . In the case of an exchangeable correlation structure, ρ_{ij} are identically equal to a scalar ρ . In the case of an auto-regressive with degree one correlation structure, ρ_{jk} are functions of a scalar parameter ρ . In more general correlation structures, such as the un-structured correlation structure, ρ_{ij} are functions of some vector-valued parameter ρ .

For the estimating equation in (5), any plug-in value of ρ between 0 and 1 will give a consistent estimate of β . When the true value ρ_0 or a consistent estimate of ρ_0 is plugged into (5), the estimate of β is consistent and efficient within a class of linear estimating equations, as long as the marginal model is correct, according to McCulloch et al. (2008). An estimator of ρ is given in the next subsection, and theorems justifying the above remarks will be given in Subsection 3.4.

3.2 Estimating ρ via composite likelihood.

Concerned with the computation burden discussed before, we choose to maximize a composite likelihood function over ρ with a fixed β . The composite likelihood for a single cluster is just the summation of all pairwise likelihoods. Denote

$$\begin{aligned} p_{ij} &= \text{pr}(Y_{ij} = 1 \mid X_{ij} = x_{ij}) = g(x_{ij}^T \beta), \\ p_{ijk} &= \text{pr}(Y_{ij} = Y_{ik} = 1 \mid X_{ij} = x_{ij}, X_{ik} = x_{ik}) = \left[(1 - \rho_{jk})e^{-(x_{ij} + x_{ik})^T \beta} + e^{-x_{ij}^T \beta} + e^{-x_{ik}^T \beta} + 1 \right]^{-1}. \end{aligned}$$

For a dataset containing m independent clusters, the composite log-likelihood is defined as

$$\begin{aligned} & \frac{1}{m} \sum_{i=1}^m \sum_{j < k} l_{jk}(X_i, Y_i; \theta) \\ &= \frac{1}{m} \sum_{i=1}^m \sum_{j < k} \left(y_{ij} y_{ik} \log p_{ijk} + (1 - y_{ij}) y_{ik} \log(p_{ik} - p_{ijk}) \right. \\ & \quad \left. + y_{ij} (1 - y_{ik}) \log(p_{ij} - p_{ijk}) + (1 - y_{ij})(1 - y_{ik}) \log(1 - p_{ij} - p_{ik} + p_{ijk}) \right). \end{aligned}$$

To estimate ρ , we solve for the equation

$$\frac{1}{m} \sum_{i=1}^m \sum_{j < k} \frac{\partial l_{jk}}{\partial \rho}(X_i, Y_i; \theta) = 0. \quad (6)$$

To estimate β and ρ jointly, Kuk (2007) suggested alternating between solving (5) with a fixed plug-in ρ from (6), and solving (6) with a fixed plug-in β from (5), until convergence, obtaining estimates $(\hat{\beta}_m, \hat{\rho}_m)$. We write m to indicate an estimate based on a dataset containing m independent clusters. This method can be viewed as a generalization of alternating logistic regression proposed by Carey et al. (1993).

3.3 Simplification of estimation procedure

We can reduce the above alternating estimation procedure of (β, ρ) into four steps:

Step 1. Solving (5) with a fixed parameter ρ_1 , obtaining $\hat{\beta}_{1m}$;

Step 2. Solving (6) with the plug-in $\hat{\beta}_{1m}$, obtaining $\hat{\rho}_{2m}$;

Step 3. Solving (5) with the plug-in $\hat{\rho}_{2m}$, obtaining $\hat{\beta}_{2m}$;

Step 4. Solving (6) with plug-in $\hat{\beta}_{2m}$, obtaining $\hat{\rho}_{3m}$.

The final estimate is $(\hat{\beta}_{2m}, \hat{\rho}_{3m})$. This simplified procedure gives an asymptotically equivalent estimate of θ as the alternating solution of (5) and (6), under a correct model specification. In the following we give a heuristic justification. A detailed proof is given in the first author's Ph.D dissertation (Zhang, 2014).

In Step 1, $\hat{\beta}_{1m}$ is a consistent estimator for β_0 , due to the robustness of (5); yet it is not efficient since ρ_1 is not necessarily the true value ρ_0 , nor a consistent estimate of ρ_0 . With a consistent estimator of β plugged into (6), $\hat{\rho}_{2m}$ is a consistent estimate of ρ_0 in Step 2. Then $\hat{\beta}_{2m}$ in Step 3 is a consistent and efficient estimate for β_0 , and $\hat{\rho}_{3m}$ in Step 4 is a consistent estimate for ρ_0 and is asymptotically equivalent to the joint solution of (5) and (6).

3.4 Large sample properties

In this section we provide several theories for the asymptotic behaviour of our estimator $(\hat{\beta}_m, \hat{\rho}_m)$.

Theorem 3.1. *Suppose conditions C1 ~ C6 stated in the appendix are satisfied, then when $m \rightarrow \infty$,*

(a) *the solution $\hat{\theta}_m = (\hat{\beta}_m, \hat{\rho}_m)$ of equations in (5) and (6) is consistent for θ_0 ;*

(b) *$\sqrt{m} \left\{ (\hat{\beta}_m - \beta_0)^T, (\hat{\rho}_m - \rho_0)^T \right\}^T$ converges weakly to a normal distribution of mean zero and a covariance matrix V given by*

$$V = \{E(B)\}^{-1} \{E(C)\} \{E(B)^T\}^{-1},$$

where

$$B = \begin{pmatrix} D(X; \beta_0)^T V^{-1}(X; \theta_0) D(X; \beta_0) & 0 \\ -\sum_{j < k} \frac{\partial^2 l_{jk}}{\partial \beta \partial \rho}(X, Y; \theta) |_{\theta_0} & -\sum_{j < k} \frac{\partial^2 l_{jk}}{\partial \rho^2}(X, Y; \theta) |_{\theta_0} \end{pmatrix},$$

$$C = \begin{pmatrix} D(X; \beta_0)^T V^{-1}(X; \theta_0) S(X, Y; \beta_0) \\ \sum_{j < k} \frac{\partial l_{jk}}{\partial \rho}(X, Y; \theta) |_{\theta_0} \end{pmatrix}^{\otimes 2}.$$

Its proof can be found in the appendix.

The next theorem is for a misspecified conditional mean model or a misspecified random effect distribution but a correct marginal mean model.

Theorem 3.2. *Suppose only the marginal mean model (4) is true, and all the other conditions in Theorem 1 are satisfied, then when $m \rightarrow \infty$,*

(a) *the solution $\hat{\theta}_m = (\hat{\beta}_m, \hat{\rho}_m)$ of equations (5) and (6) is consistent for (β_0, ρ_1) , where ρ_1 is the value that minimizing a Kullback-Leibler distance defined on composite likelihoods between the misspecified pairwise joint model and the true pairwise joint model:*

$$KL_{\text{composite}}(L, L^*) = E_0 \left[\log \left\{ \frac{\prod_{j < k} L(X_j, X_k, Y_j, Y_k; \beta_0, \eta)}{\prod_{j < k} L^*(X_j, X_k, Y_j, Y_k; \beta_0, \rho_1)} \right\} \right],$$

where L denotes the likelihood of the true pairwise joint model, L^* for the mis-specified one, and η is some other parameters under the true model.

(b) $\sqrt{m} \left\{ (\hat{\beta}_m - \beta_0)^T, (\hat{\rho}_m - \rho_1)^T \right\}^T$ converges weakly to a normal distribution of mean zero and a covariance matrix W given by

$$W = \{E(B_1)\}^{-1} \{E(C_1)\} \{E(B_1)^T\}^{-1},$$

where

$$B_1 = \begin{pmatrix} D(X; \beta_0)^T V^{-1}(X; \beta_0, \rho_1) D(X; \beta_0) & 0 \\ -\sum_{j < k} \frac{\partial^2 l_{jk}^*}{\partial \beta \partial \rho}(X, Y; \theta) |_{(\beta_0, \rho_1)} & -\sum_{j < k} \frac{\partial^2 l_{jk}^*}{\partial \rho^2}(X, Y; \theta) |_{(\beta_0, \rho_1)} \end{pmatrix},$$

$$C_1 = \begin{pmatrix} D(X; \beta_0)^T V^{-1}(X; \beta_0, \rho_1) S(X, Y; \beta_0) \\ \sum_{j < k} \frac{\partial l_{jk}^*}{\partial \rho}(X, Y; \theta) |_{(\beta_0, \rho_1)} \end{pmatrix}^{\otimes 2}.$$

As suggested in Theorem 1, when the pairwise conditional model is correct, the asymptotic covariance of $\sqrt{m}(\hat{\beta}_m - \beta_0)$ can be estimated by

$$\hat{V}_m^\beta := m \left(\sum_{i=1}^m D(X_i; \hat{\beta}_m)^T V^{-1}(X_i; \hat{\theta}_m) D(X_i; \hat{\beta}_m) \right)^{-1}.$$

Allowing for a potentially mis-specified conditional model, a robust estimate of the asymptotic covariance of $\sqrt{m}(\hat{\beta}_m - \beta_0)$ is

$$\hat{V}_m^{\text{robust}} := m \left(\sum_{i=1}^m D(X_i; \hat{\beta}_m)^T V^{-1}(X_i; \hat{\theta}_m) D(X_i; \hat{\beta}_m) \right)^{-1} \left(\sum_{i=1}^m \left[D(X_i; \hat{\beta}_m)^T V^{-1}(X_i; \hat{\theta}_m) S(X_i, Y_i; \hat{\beta}_m) \right]^{\otimes 2} \right) \\ \cdot \left(\sum_{i=1}^m D(X_i; \hat{\beta}_m)^T V^{-1}(X_i; \hat{\theta}_m) D(X_i; \hat{\beta}_m) \right)^{-1}.$$

3.5 Discussion of inference methods and further remarks

Inference by estimating equations (5) and (6) reduces the computation burden to n_i^2 for every cluster, compared to the order of 2^{n_i} in maximum likelihood inference. Alternative inference procedures may be adopted for the estimation of ρ ; an example is

the second-order GEE in Prentice (1988). However, the computational burden of that method is in the order of $O(n_i^6)$, since it computes the inverse of a $n_i^2 \times n_i^2$ matrix, which is the weighting matrix for pairwise outcome products.

Under misspecification of the conditional distribution or the random effect distribution, the estimating equation (5) still guarantees consistency of the marginal parameter β , while the inverse weighting matrix V is still a genuine covariance matrix, but corresponds to a misspecified model.

4 Generalization to three-level clustered data

For notational simplicity, our earlier discussions focused on two-level clustered data. Since our proposed model allows for flexible modeling of correlations between individual observations similar to Gaussian mixed effect models, it can be readily extended to datasets with a higher level of clustering. In this section, we consider a three-level clustered data where the first level consists of multiple independent clusters, inside each nested multiple individuals representing the second level, and multiple observations taken on every individual from the third level. Observations from different clusters are independent. Data from the i^{th} cluster can be denoted by $(X_i, Y_i) = \text{vec}(X_{ijk}, Y_{ijk}) : j = 1, \dots, n_i$ indexes individuals from the i^{th} cluster and $k = 1, \dots, n_{ij}$ counts observations on the j^{th} individual from the i^{th} cluster.

We assume a similar working conditional model:

$$\text{pr}(Y_{ijk} = 1 \mid X_{ijk}, a_{ijk}) = \exp\left(-a_{ijk}e^{-X_{ijk}^T\beta}\right), \quad a_{ijk} \sim \text{Exp}(1).$$

It is easy to show that the marginalization property of the working model still holds in the case of three-level clustering data:

$$\text{pr}(Y_{ijk} = 1 \mid X_{ijk} = x_{ijk}) = \frac{e^{x_{ijk}^T\beta}}{1 + e^{x_{ijk}^T\beta}}.$$

One way to model correlations among a_{ijk} 's is to assume that the level-two observations are exchangeable, and the level-three observations nested within level-two are also exchangeable. To be specific, we can model the correlations as follows:

$$\text{cor}(a_{ijk}, a_{ij'k'}) = \rho_2, \quad j \neq j' \tag{7}$$

$$\text{cor}(a_{ijk}, a_{ij'k'}) = \rho_2 + \rho_3, \quad j = j', \quad k \neq k'. \tag{8}$$

Similar robust estimation methods based on (5) and (6) can still be used in this case. Denote $N_i = \sum_{j=1}^{n_i} n_{ij}$ being the total number of observations from cluster i . For notational simplicity, we concatenate level-two observations in the cluster and denote $(X_i, Y_i) = \{\text{vec}(X_s, Y_s) : s = 1, \dots, N_i\}$; i.e., we merge the double index jk into a single index s . Suppose distinct observations s_1, s_2 are from individuals j_1, j_2 in the i^{th} cluster respectively, then $\sum_{j=1}^{j_1-1} n_{ij} < s_1 \leq \sum_{j=1}^{j_2} n_{ij}$, $l = 1, 2$.

Entries of the covariance matrix $V(X_i, \beta, \rho)$ are given by:

$$V_{s_1 s_1}(X_i; \beta, \rho) = \frac{e^{-X_{is_1}^T \beta}}{\left(1 + e^{-X_{is_1}^T \beta}\right)^2},$$

$$V_{s_1 s_2}(X_i; \beta, \rho) = \frac{1}{\{1 - \text{cor}(a_{is_1}, a_{is_2})\} e^{-(X_{is_1} + X_{is_2})^T \beta} + e^{-X_{is_1}^T \beta} + e^{-X_{is_2}^T \beta} + 1}} - \frac{1}{1 + e^{-X_{is_1}^T \beta}} \frac{1}{1 + e^{-X_{is_2}^T \beta}}.$$

If we follow the exchangeable correlation formulation in (7) and (8),

$$V_{s_1 s_2}(X_i; \beta, \rho) = \begin{cases} \left\{ (1 - \rho_2 - \rho_3) e^{-(X_{is_1} + X_{is_2})^T \beta} + e^{-X_{is_1}^T \beta} + e^{-X_{is_2}^T \beta} + 1 \right\}^{-1} - \left\{ (e^{-X_{is_1}^T \beta} + 1) (e^{-X_{is_2}^T \beta} + 1) \right\}^{-1}, & j_1 = j_2 \\ \left\{ (1 - \rho_2) e^{-(X_{is_1} + X_{is_2})^T \beta} + e^{-X_{is_1}^T \beta} + e^{-X_{is_2}^T \beta} + 1 \right\}^{-1} - \left\{ (e^{-X_{is_1}^T \beta} + 1) (e^{-X_{is_2}^T \beta} + 1) \right\}^{-1}, & j_1 \neq j_2. \end{cases}$$

Similar to (6), we write

$$\begin{aligned} & \sum_{i=1}^m \sum_{s_1 < s_2} l_{s_1 s_2}(X_i, Y_i, \theta) \\ &= \sum_{i=1}^m \sum_{s_1 < s_2} \{ y_{is_1} y_{is_2} \log p_{is_1 s_2} + (1 - y_{is_1}) y_{is_2} \log(p_{is_2} - p_{is_1 s_2}) \\ & \quad + y_{is_1} (1 - y_{is_2}) \log(p_{is_1} - p_{is_1 s_2}) + (1 - y_{is_1}) (1 - y_{is_2}) \log(1 - p_{is_1} - p_{is_2} + p_{is_1 s_2}) \}, \end{aligned}$$

where $p_{is_1} = \left(1 + e^{-X_{is_1}^T \beta}\right)^{-1}$ and

$$p_{is_1 s_2} = \begin{cases} \left\{ (1 - \rho_2 - \rho_3) e^{-(X_{is_1} + X_{is_2})^T \beta} + e^{-X_{is_1}^T \beta} + e^{-X_{is_2}^T \beta} + 1 \right\}^{-1}, & j_1 = j_2, \\ \left\{ (1 - \rho_2) e^{-(X_{is_1} + X_{is_2})^T \beta} + e^{-X_{is_1}^T \beta} + e^{-X_{is_2}^T \beta} + 1 \right\}^{-1}, & j_1 \neq j_2. \end{cases}$$

Similar to the case of two-level clustering, estimates are obtained by solving

$$\begin{cases} \frac{1}{m} \sum_{i=1}^m D(X_i; \beta) V^{-1}(X_i; \beta, \rho) S(X_i, Y_i; \beta) = 0, \\ \frac{1}{m} \sum_{i=1}^m \sum_{s_1 < s_2} \frac{\partial l_{s_1 s_2}}{\partial \rho}(X_i, Y_i; \beta, \rho) = 0. \end{cases}$$

Other correlation structures can also be used. For example, suppose the level-two observations are exchangeable units and the level-three observations are auto-regressive with order one, then we could model

$$\begin{aligned} \text{cor}(a_{is_1}, a_{is_2}) &= \rho_2, \quad j_1 \neq j_2, \\ \text{cor}(a_{is_1}, a_{is_2}) &= \rho_2 + \rho_3^{|s_1 - s_2|}, \quad j_1 = j_2. \end{aligned}$$

Entries in the inverse weighting matrix for estimating β can be written as

$$V_{s_1 s_2}(X_i; \beta, \rho) = \frac{1}{\{1 - \text{cor}(a_{is_1}, a_{is_2})\} e^{-(X_{is_1} + X_{is_2})^T \beta} + e^{-X_{is_1}^T \beta} + e^{-X_{is_2}^T \beta} + 1} - \frac{1}{1 + e^{-X_{is_1}^T \beta}} \frac{1}{1 + e^{-X_{is_2}^T \beta}}.$$

We can write

$$V_{s_1 s_2}(X_i; \beta, \rho) = \begin{cases} \left\{ (1 - \rho_2 - \rho_3^{|s_1 - s_2|}) e^{-(X_{is_1} + X_{is_2})^T \beta} + e^{-X_{is_1}^T \beta} + e^{-X_{is_2}^T \beta} + 1 \right\}^{-1} - \left\{ (e^{-X_{is_1}^T \beta} + 1) (e^{-X_{is_2}^T \beta} + 1) \right\}^{-1}, & j_1 = j_2, \\ \left\{ (1 - \rho_2) e^{-(X_{is_1} + X_{is_2})^T \beta} + e^{-X_{is_1}^T \beta} + e^{-X_{is_2}^T \beta} + 1 \right\}^{-1} - \left\{ (e^{-X_{is_1}^T \beta} + 1) (e^{-X_{is_2}^T \beta} + 1) \right\}^{-1}, & j_1 \neq j_2. \end{cases}$$

The four-step iterative estimation in Subsection 3.3 still applies to this setting.

5 Numerical Studies

5.1 Simulation

We conducted simulation studies to evaluate the finite sample performance of our proposed estimators. In each simulation scenario, 1000 Monte Carlo datasets were generated. In each dataset, we generated 200 independent clusters. A covariate X_1 is included, which was a continuous normal random variable with mean zero and standard deviation 2.

Throughout this subsection, the marginal model was assumed to be

$$\text{pr}(Y_{ij} = 1 \mid X_{ij}) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 X_{ij1})}, \quad (9)$$

where $\beta_0 = 1$ and $\beta_1 = -1.2$. Under scenarios with joint distributions complying to our proposed models, we generated frailties from a multivariate standard exponential distribution with varying correlation structures, by the procedure discussed in Subsection 2.4. To be specific, for the cases of two-level clustering, in which cluster sizes varied from 5 to 7 with equal probabilities, we imposed an exchangeable correlation structure and an auto-regressive of order one correlation structure. Exchangeable correlation structure is typically implemented to model correlations between individuals sampled from the same geographical region, hospital, etc; auto-regressive correlation usually models longitudinal observations over time. For the case of three-level clustering, we put 2 or 3 individuals into each cluster with probabilities 4/5 and 1/5 and generated 2 or 3 observations for each individual with probabilities 4/5 and 1/5. We imposed exchangeable correlation structures for both levels of clustering as discussed in Section 4. For model inference, we assumed the correct joint model and only model-based standard errors and 95% confidence interval coverage rates were listed since their robust counterparts behaved quite similarly.

A misspecified joint model was also considered, in which correlation was introduced via a latent variable model. For each cluster i , we generated an uniform variable U_i and transformed it into a logistic distributed random variable $A_i = \log U_i - \log(1 - U_i)$; at the end, we simulated $(Y_{i1}, \dots, Y_{in_i})$ by $Y_{ij} = I(X_{ij}^T \beta + A_i > 0)$, which satisfies the marginal model in (9). For the

proposed inference method, both the model-based and the robust standard errors and their respective 95% confidence interval coverage rates were presented.

Table 1 lists the simulation results in the case of two-level clustering under an exchangeable correlation structure and an auto-regressive of order one correlation structure, respectively in (a) and (b), under correctly specified joint models. The estimation efficiencies of our estimates, measured by mean squared error (MSE), are quite close to the MLE's, but the proposed method takes much less computing time than MLE. When $\rho = 0.9$, β estimate from MLE has a much larger bias compared to the proposed inference method.

Table 2 lists simulation results for three-level clustering. When the correlation is small, results from the two inference methods are pretty close. Otherwise, MLE estimates of β are more biased. Besides, MLE behaves much worse than the proposed method in estimating the correlation parameters even when the correlation level is mild.

Table 3 lists simulation results for the mis-specified conditional model case. As expected, MLE of β is biased while the proposed method gives consistent estimates of β , along with consistently estimated robust standard errors.

5.2 Madras longitudinal schizophrenia study

We further demonstrate our proposed method using the Madras longitudinal schizophrenia study from Thara et al. (1994), in which first-episode schizophrenics were followed for 10 years with the primary objective of characterizing the natural history of disease progression. The data contain several longitudinal binary outcome measurements indicating the presence of positive psychiatric symptoms over the time course: $t_{ij} = 0, \dots, 11$ months during the first year following an initial hospitalization for 86 schizophrenia patients. The binary outcome Y_{ij} under interest is an indicator of whether or not a patient is observed to have thought disorders. Covariates include the time variable t_{ij} , a binary indicator X_{ij2} of whether or not a patient is younger than 20 at disease onset and gender X_{ij3} : 0 for male and 1 for female. To assess the association between occurrence of thought disorders and the covariates, a marginal logistic regression model is constructed using a linear trend in time, with the time-independent binary covariates X_{ij2} and X_{ij3} :

$$\text{logit}E(Y_{ij} | t_{ij}, X_{ij}) = \beta_0 + \beta_1 t_{ij} + \beta_2 X_{ij2} + \beta_3 X_{ij3}.$$

Our regression model is almost identical to the model from Heagerty (1999), except that we did not center the time covariate.

Using the proposed method, we can answer whether the population-averaged probability of thought disorders differs across time, age-at-onset and gender subgroups. We used our proposed method and maximum likelihood to analyze this dataset, assuming observations from the same patients are exchangeable and auto-regressive with order one over time, i.e. AR(1). The results are reported in Table 4. We can see the results from different inference methods are pretty similar, and the length of 95% confidence intervals based on the proposed method is similar to those based on MLE. Since this is a longitudinal dataset, auto-regressive of order one (in time unit) correlation structure should be more close to the real situation, and in the following we report the results from our proposed inference method.

The estimated odds of thought disorder prevalence for a patient younger than 20 at the beginning of hospitalization is 47%

higher (95% C.I.: 19% lower to 166% higher) than elder patient, controlling for gender and observation time. The estimated odds of thought disorder prevalence for a female patient is 46% lower (95% CI: 70% lower to 4% lower) than a male patient, controlling for age at onset and observation time. The estimated odds of thought disorder decreases by 29% (95% CI: 33% to 24%) in one month during hospitalization, controlling for age at onset and gender. There is evidence of significant decrease in thought disorder occurrence probability as times passes in hospital; or comparing females to males.

5.3 British Social Attitudes Panel Survey

To demonstrate our method for three-level clustered data, we analyzed the *British Social Attitudes Panel Survey* conducted from 1983 through 1986. In this survey, subjects were asked whether they thought there should be no legal or governmental regulation on abortion. This survey was carried out in 54 districts annually for four years among the same individuals. The dataset includes people who have completed all four surveys during the four years, adding up to 1,056 observations from 264 individuals in total. Covariates can be categorized into three levels: the first level is a district-level covariate: the percent of protestants of each district; the second level includes individual-level demographic covariates, including social class (middle, upper and lower), gender (male and female) and religion (Protestant, Catholic, other and none); in the third levels are three dummy variables for years 1984, 1985, 1986. We can see there are two covariates corresponding to protestant in the model, one on the district-level and the other on the individual-level. By this arrangement we are able to estimate the effect of protestant religion both within district and between districts, as discussed in Neuhaus and Kalbfleisch (1998). The inclusion of the two protestant variables are potentially of substantive interest by measuring the religious context or environment impact on individual attitude in contrast to their own religious affiliation affect, as discussed in Heagerty and Zeger (2000).

In Table 5, point estimates and 95% confidence intervals of odds ratio corresponding to the above covariates are listed, from three methods. Method 1 is our proposed method assuming the correlation structure of random effects within individuals is auto-regressive of order one and the correlation structure across individuals within a district is exchangeable. Method 2 is also carried out by our proposed method, but assuming both correlation structures within districts and individuals are exchangeable. Method 3 is GEE with an exchangeable working correlation matrix for observations within a district. This ignores the finer level of correlation between observations within individuals, by assuming correlations being equal both within an individual and between two individuals from the same district. We did not compare the results with the MLE as the algorithm failed to converge.

Method 1 and Method 2 give out roughly the same point estimates as GEE but with narrower 95% confidence intervals. The exceptions are categorical covariates representing religion contrasts between other religions and Protestants within districts having similar proportions of Protestants. This can be explained by the relatively small sample size of this subgroup. In total, we only have 45 individuals of other religions.

Comparing Methods 1 and 2, we can see that the results are roughly the same, indicating the robustness of the proposed method with respect to different assumed correlation structures. Since Method 1 assumes an auto-regressive correlation structure on the third level, where observations are taken annually and therefore their correlations can be better described by an auto-regressive correlation structure. In the following we report results from Method 1.

The covariates we put into Method 1 decompose religion contrasts into within-cluster contrast and between-cluster contrast. The variable %Protestant is a district-level covariate and equals to the sampled proportion of district Protestants. The estimated odds ratio is 2.17, 95% CI: (0.86, 5.52), indicating a non-significantly increasing trend of allowing abortions among individuals from districts of a higher level in Protestants, controlling for all the other variables. The categorical religion contrasts Catholic, other, none, to the reference Protestant group can then be interpreted as comparing the propensity of allowing abortion among individuals of different religions who reside in districts of equal level in Protestants, controlling for year surveyed, social class and gender. Non-significantly lower odds are observed among Catholics in contrast to Protestants with a ratio of 0.67, 95% CI: (0.25, 1.80), non-significantly lower odds are observed among those of other religions with the ratio 0.52, 95% CI: (0.24, 1.11) and significantly higher odds are observed among those without any religions with odds ratio being 2.00, 95% CI: (1.21, 3.30). The propensity of allowing abortions among females is non-significantly lower than that among males from districts of equal level in Protestants, controlling for working class, religion and year of survey, with an odds ratio as 0.72 95% CI: (0.48, 1.07). The odds ratio of allowing abortions from upper working class comparing to middle class is 0.76, 95% CI: (0.51, 1.14), and the odds ratio comparing lower working class to middle class is 0.80, 95% CI: (0.54, 1.19), among people from districts of equal level in Protestants, controlling for gender, religion and year of survey. As for time trend in allowing for abortions propensity, there is a significant drop in Year 1984 compared to the previous year with odds 0.66, 95% CI: (0.49, 0.88), and there are non-significant increments in the following two years, compared to Year 1983.

6 Concluding remarks

In this paper we introduce a marginalizable conditional model for analysing clustered binary data. A working generalized linear mixed effect model and a multivariate Gumbel random intercept distribution are proposed, which yield a marginal logistic regression model that has a population-level interpretation.

Unlike most marginal models which model the first and perhaps the second moment, we have come up with a parametric marginal model, which guarantees there is always a real joint distribution for the marginal logistic regression model and parameters being estimated always exist. In contrast, one criticism of GEE with a cluster-common working correlation matrix for a binary outcome is that there may not be any multivariate distribution with a correlation structure being equivalent to GEE's working correlation structure.

By generalizing the estimating equation from alternating logistic regression proposed by Carey et al. (1993), our proposed inference yields consistent estimates of marginal parameters even under misspecified conditional model or random effect distribution, along with consistent estimates of estimators standard deviation.

The marginalization property is based on a standard exponential frailty assumption, which can be viewed as a special case of the Gamma frailty models considered in Henderson and Shimakura (2003) and Coull et al. (2006). However for more general Gamma distributions, the marginal model is no longer conveniently interpretable. Exponential distributed frailties should not be considered as a limitation, since

1. a marginal logistic model interpretation is often desirable in practice;

2. an exponential distributed frailty is equivalent to a Gumbel random intercept which has physical interpretations. Gumbel distribution can model the distribution of maximum of the normal or exponential type random variables, so Gumbel random intercept is reasonable when we believe there are many latent cluster effects and the maximum dominates the others; i.e. the random effect can be modeled as the maximum of many cluster effects;
3. robust estimation procedure being proposed would yield consistent estimates for marginal parameters even when the multivariate exponential frailty distribution or the conditional mean model is misspecified.
4. marginal inference is un-affected when frailty distribution is covariate dependent.

In this paper we have concentrated on correlated binary outcomes. In principle, our model can be generalized into the cases of correlated ordinal and censored survival data. Investigations are being carried on along these directions.

Acknowledgment

The authors thank Professor Jon A. Wellner for his helpful comments on the proof of the theoretical results.

Appendix

Here we list conditions of Theorem 1 and prove it. The proof of Theorem 2 is very similar and is omitted.

$$\begin{aligned} \text{We define } \Psi(\theta) &= \begin{pmatrix} \Psi_1(\theta) \\ \Psi_2(\theta) \end{pmatrix} = \begin{pmatrix} E\{f_1(X, Y; \theta)\} \\ E\{f_2(X, Y; \theta)\} \end{pmatrix}, \\ \text{and } \Psi_m(\theta) &= \begin{pmatrix} \Psi_{1,m}(\theta) \\ \Psi_{2,m}(\theta) \end{pmatrix} = \begin{pmatrix} \frac{1}{m} \sum_{i=1}^m f_1(X_i, Y_i; \theta) \\ \frac{1}{m} \sum_{i=1}^m f_2(X_i, Y_i; \theta) \end{pmatrix}, \end{aligned}$$

where f_1 and f_2 correspond to estimating equations in (5) and (6):

$$\begin{cases} f_1(X_i, Y_i; \theta) := D(X_i; \beta)^T V^{-1}(X_i; \theta) S(X_i, Y_i; \beta), \\ f_2(X_i, Y_i; \theta) := \sum_{j < k} \frac{\partial l_{jk}}{\partial \rho}(X_i, Y_i; \theta). \end{cases}$$

Theorem 1 is true under the following conditions:

- C.1 Observations from different clusters are independent and identically distributed.
- C.2 Number of observations per cluster is uniformly bounded.
- C.3 Parameter space Θ is a convex and compact subset of \mathbb{R}^p and the true value of parameter, θ_0 , is not a boundary point of Θ .
- C.4 The probability of covariate X being degenerate is 0, i.e., $X^T \beta = 0$ a.e. implies $\beta = 0$ a.e., and X is bounded with probability one.

C.5 There is a unique root of β from $\Psi_1(\beta, \rho) = 0$ for all ρ .

C.6 The joint distribution is correctly specified.

Proof. First we would like to point out that even though different clusters may contain different numbers of observations, we can still view the joint observations from a cluster as independent and identically distributed (i.i.d).

We can regard each cluster in theory contains infinite subjects and their quantities are denoted by $(X(\cdot), Y(\cdot), a(\cdot))$: \cdot varies with different subjects. The data we observe from a cluster is a deterministic projection of $(X(\cdot), Y(\cdot), a(\cdot))$. Assuming the stochastic process $(X(\cdot), Y(\cdot), a(\cdot))$ are i.i.d. and the projection procedure is also i.i.d., we conclude observations from different clusters are i.i.d.. We denote P_0 as the joint distribution.

Second, we would like to argue that ρ_0 is the unique solution to $\Psi_2(\theta) = 0$ at $\beta = \beta_0$. This can be shown by the Kullback-Leibler divergence for composite likelihood.

Composite likelihood of the i^{th} cluster is

$$\prod_{j < k} L_{jk}(X_i, Y_i; \theta)$$

The Kullback-Leibler divergence for composite likelihood is

$$KL_{\text{composite}}(L_0, L_1) := P_0 \log \left(\frac{\prod_{j < k} L_0(X_j, X_k, Y_j, Y_k; \beta_0, \rho_0)}{\prod_{j < k} L_1(X_j, X_k, Y_j, Y_k; \beta_0, \rho_1)} \right) = \sum_{j < k} P_0 \log \left(\frac{L_{jk}(X_i, Y_i; \beta_0, \rho_0)}{L_{jk}(X_i, Y_i; \beta_0, \rho_1)} \right) > 0$$

the last strict in-equality is due to Jensen's Inequality and the fact that $L_1 = L_0$ if and only if $\rho_1 = \rho_0$.

Thus ρ_0 is the unique value maximizing composite likelihood expectation with plug-in β_0 . Since the model is smooth in parameters, $\Psi_2(\theta) = 0$ uniquely at $\rho = \rho_0$ when β is fixed at β_0 .

Next, consider an index set $\mathcal{H} := \{h \in \mathbb{R}^p : \|h\| \leq 1\}$ in which $\|\cdot\|$ is the Euclidean norm. Then the following function class indexed by $\theta \in \Theta$ and $h \in \mathcal{H}$, defined on the sample space of (X, Y) , i.e. cluster observations:

$$\mathcal{F}_0 := \{h^T(f_1(X, Y; \theta), f_2(X, Y; \theta)) : \theta \in \Theta, h \in \mathcal{H}, (X, Y) \sim P_0\}$$

is P_0 -Donsker.

For an arbitrary pair of functions from \mathcal{F}_0 :

$$\begin{aligned} & |h_1^T(f_1(X, Y; \theta_1), f_2(X, Y; \theta_1)) - h_2^T(f_1(X, Y; \theta_2), f_2(X, Y; \theta_2))| \\ & \leq C_0 \|\theta_1 - \theta_2\| \cdot \|h_1 - h_2\| \end{aligned} \tag{10}$$

This is due to the fact that everything in $h^T(f_1(X, Y; \theta), f_2(X, Y; \theta))$ is continuous in θ so Mean Value Theorem can be used based on conditions C.2 and C.3; C_0 is some finite number by condition C.4.

Since $\theta_1, \theta_2 \in \Theta$ and Θ is a compact subset of Euclidean space, number of brackets needed to cover \mathcal{F}_0 satisfies P_0 -Donsker requirement, according to van der Vaart and Wellner (1996), page 129.

Now we can claim

$$\sup_{\theta \in \Theta, h \in \mathcal{H}} |h^T \Psi_m(\theta) - h^T \Psi(\theta)| \rightarrow 0$$

implying that

$$\begin{aligned} & \sup_{h \in \mathcal{H}} |h^T [\Psi_m(\hat{\theta}_m) - \Psi(\hat{\theta}_m)]| \rightarrow 0, \\ \text{i.e.} \quad & \sup_{h \in \mathcal{H}} |h^T \Psi(\hat{\theta}_m)| \rightarrow 0; \quad \text{thus, } |\Psi(\hat{\theta}_m)| \rightarrow 0. \end{aligned}$$

Since $(f_1(X, Y; \theta), f_2(X, Y; \theta))$ are continuous in θ , we have shown $\hat{\theta}_m \xrightarrow{P} \theta_0$. \square

The proof of the weak convergence of $\sqrt{m} \left\{ (\hat{\beta}_m - \beta_0)^T, (\hat{\rho}_m - \rho_0)^T \right\}^T$ makes use of Theorem 3.3.1 of van der Vaart and Wellner (1996), which is stated as the following.

Suppose there are two random mappings Ψ_m and Ψ such that $\Psi(\beta_0, \rho_0) = 0$ for some interior point $(\beta_0, \rho_0) \in \Theta$, $\Psi_m(\beta_m, \rho_m) \xrightarrow{P} 0$ for some random sequence $(\beta_m, \rho_m) \subset \Theta$, and assume the followings are true:

P.1 (β_m, ρ_m) is consistent for (β_0, ρ_0) ;

P.2 $\sqrt{m} (\Psi_m - \Psi) (\beta_0, \rho_0)$ converges in distribution to a tight random element Z ;

P.3

$$\begin{aligned} & \sqrt{m} (\Psi_m - \Psi) (\beta_m, \rho_m) - \sqrt{m} (\Psi_m - \Psi) (\beta_0, \rho_0) \\ = & o_p \left(1 + \sqrt{m} \|\beta_m - \beta_0\| + \sqrt{m} \|\rho_m - \rho_0\| \right); \end{aligned}$$

P.4 $\Psi(\beta, \rho)$ is Fréchet differentiable at (β_0, ρ_0) ;

P.5 The derivative of $\Psi(\beta, \rho)$ at (β_0, ρ_0) , denoted by $\dot{\Psi}(\beta_0, \rho_0)$, is continuously invertible.

Then

$$\sqrt{m} \left\{ (\hat{\beta}_m - \beta_0)^T, (\hat{\rho}_m - \rho_0)^T \right\}^T \xrightarrow{d} -\dot{\Psi}(\beta_0, \rho_0)^{-1}(Z).$$

Proof. Condition P.1 has been verified.

Since we have shown \mathcal{F}_0 is P_0 -Donsker, condition P.2 is verified.

By P_0 -Donsker preservation theorem 2.10.3 in van der Vaart and Wellner (1996), this function class

$$\left\{ h^T [(f_1(\beta, \rho), f_2(\beta, \rho)) - (f_1(\beta_0, \rho_0), f_2(\beta_0, \rho_0))] : (\beta, \rho) \in \Theta, h \in \mathcal{H} \right\}$$

is P_0 -Donsker as well.

$$\begin{aligned} & \sup_{h \in \mathcal{H}} P_0 \left(h^T [(f_1(\beta, \rho), f_2(\beta, \rho)) - (f_1(\beta_0, \rho_0), f_2(\beta_0, \rho_0))] \right)^2 \\ \leq & P_0 (C_0 \|\theta_0 - \theta\|)^2 \rightarrow 0 \quad \text{as } \|(\beta, \rho) - (\beta_0, \rho_0)\| \rightarrow 0 \end{aligned}$$

Therefore, according to Lemma 3.3.5 of van der Vaart and Wellner (1996), P.3 holds.

As for P.4, since Ψ is a smooth function in parameters, it is trivial to verify that $-E(B)$ is its Fréchet derivative at (β_0, ρ_0) . Due to model identifiability and condition C.3, $E(B)$ is a negative definite matrix and thus continuously invertible. Therefore, P.5 is also satisfied. \square

References

- Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):pp. 9–25.
- Carey, V., Zeger, S. L., and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, 80(3):517–526.
- Chaganty, N. R. and Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):851–860.
- Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):pp. 141–151.
- Conaway, M. (1990). A random effects model for binary data. *Biometrics*, pages 317–328.
- Coull, B. A., Houseman, E. A., and Betensky, R. A. (2006). A computationally tractable multivariate random effects model for clustered binary data. *Biometrika*, 93(3):pp. 587–599.
- Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, 55(3):688–698.
- Heagerty, P. J. and Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, 15(1):pp. 1–19.
- Henderson, R. and Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika*, 90(2):pp. 355–366.
- Hougaard, P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity. *Biometrika*, 71(1):pp. 75–83.
- Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 48(3):pp. 795–806.
- Krishnamoorthy, A. S. and Parthasarathy, M. (1951). A multivariate gamma-type distribution. *The Annals of Mathematical Statistics*, 22(4):pp. 549–557.
- Kuk, A. Y. C. (2007). A hybrid pairwise likelihood method. *Biometrika*, 94(4):939–952.

- Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability.
- McCulloch, C. E., Searle, S. R., and M., N. J. (2008). *Generalized, linear, and mixed models*. Wiley.
- Neuhaus, J., Kalbfleisch, J., and Hauck, W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review/Revue Internationale de Statistique*, pages 25–35.
- Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54(2):pp. 638–645.
- Oakes, D. (1982). A model for association in bivariate survival data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(3):pp. 414–422.
- O’Brien, S. M. and Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics*, 60(3):pp. 739–746.
- Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):pp. 12–35.
- Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, 44(4):pp. 1033–1048.
- Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(4):pp. 749–760.
- Song, P. X.-K., Li, M., and Yuan, Y. (2009). Joint regression analysis of correlated data using gaussian copulas. *Biometrics*, 65(1):60–68.
- Thara, R., Henrietta, M., Joseph, A., Rajkumar, S., and Eaton, W. W. (1994). Ten-year course of schizophrenia: the madras longitudinal study. *Acta Psychiatrica Scandinavica*, 90(5):329–336.
- Vaida, F. and Xu, R. (2000). Proportional hazards model with random effects. *Statistics in Medicine*, 19(24):3309–3324.
- van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer.
- Wang, Z. and Louis, T. A. (2004). Marginalized binary mixed-effects models with covariate-dependent random effects and likelihood inference. *Biometrics*, 60(4):pp. 884–891.

Table 1: Simulation results for estimating $(\beta_0, \beta_1, \rho_0)$ in two-level clustering, where ρ_0 is the correlation parameter of random effects. Bias represents the empirical bias, SSE represents the Monte Carlo standard error (s.e.), MSE is the mean squared error. SEE represents the averaged model-based s.e. estimates.

(a) Two-level clustering, exchangeable correlation matrix.													
ρ_0	Method	Bias $\times 10^3$		SEE $\times 10^3$		SSE $\times 10^3$		MSE $\times 10^3$		95% C.I. coverage rate		Bias $\times 10^3$	Computing Times (sec)
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}_0$	
0.1	Proposed	-1	3	93	71	93	69	9	5	96.0%	96.2%	6	5
	MLE	-1	3	93	71	93	69	9	5	96.0%	96.2%	2	44
0.3	Proposed	-5	7	98	71	103	73	11	5	93.6%	95.4%	-16	6
	MLE	-5	7	98	71	103	73	11	5	93.7%	95.4%	-16	51
0.5	Proposed	-6	2	104	71	106	69	11	5	94.5%	96.0%	-13	6
	MLE	-6	2	104	71	106	69	11	5	94.5%	96.0%	-12	53
0.7	Proposed	-8	2	112	72	114	71	13	5	94.9%	95.0%	-11	6
	MLE	-9	2	112	72	113	71	13	5	94.4%	95.0%	-9	51
0.9	Proposed	-9	7	123	75	121	72	15	5	94.4%	96.6%	-9	6
	MLE	68	-47	117	69	111	63	17	6	91.4%	89.4%	71	35
(b) Two-level clustering, AR(1) correlation matrix.													
ρ_0	Method	Bias $\times 10^3$		SEE $\times 10^3$		SSE $\times 10^3$		MSE $\times 10^3$		95% C.I. coverage rate		Bias $\times 10^3$	Computing Times (sec)
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}_0$	
0.1	Proposed	-4	6	92	71	88	72	8	5	95.6%	95.1%	30	5
	MLE	-4	6	92	71	88	72	8	5.2	95.6%	95.0%	36	75
0.3	Proposed	-9	11	94	71	95	74	9	6	94.0%	94.0%	-25	7
	MLE	-9	11	94	71	95	74	9	6	94.1%	94.2%	-16	63
0.5	Proposed	-7	6	98	71	106	72	11	5	93.8%	94.6%	-20	7
	MLE	-7	66	98	71	106	72	11	5	93.6%	94.7%	-17	62
0.7	Proposed	-8	9	104	72	106	73	11	5	94.6%	94.8%	-16	9
	MLE	-9	9	104	72	106	73	11	5	94.4%	95.0%	-11	63
0.9	Proposed	-6	8	114	73	118	71	14	5	94.3%	95%	-7	31
	MLE	134	-89	111	66	130	80	35	14.3	72.3%	63.8%	-14	35

Table 2: Simulation results for estimating $(\beta_0, \beta_1, \rho_2, \rho_3)$ in a three-level clustering. We assume exchangeable correlation structure in both levels of clustering, and (ρ_2, ρ_3) represents the true correlations in the second and the third clustering levels. Bias, SSE, SEE, MSE represent the same quantities as in Table 1. 95% confidence interval coverage rates are presented, derived from model based s.e..

Three-level clustering, exchangeable correlation matrix.															
ρ_2	ρ_3	Method	Bias $\times 10^3$		SEE $\times 10^3$		SSE $\times 10^3$		MSE $\times 10^3$		95% C.I. coverage rate		Bias $\times 10^3$		Computing Times (sec)
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}_2$	$\hat{\rho}_3$	
0.1	0.1	Proposed	-4	9	84	121	85	124	7	15	94.7%	95.1%	13	53	36
		MLE	-4	10	80	121	85	124	7	16	93.6%	95.3%	-68	-61	57
0.1	0.3	Proposed	< 1	2	86	120	88	123	8	15	94.3%	95.2%	18	10	23
		MLE	< 1	2	82	121	89	124	8	15	93.2%	95.2%	-27	-184	72
0.1	0.5	Proposed	5	6	89	120	89	123	8	15	95.6%	94.8%	15	17	21
		MLE	2	-2	85	120	93	127	9	16	93.4%	94.3%	-12	-247	78
0.1	0.7	Proposed	-1	6	93	118	90	127	8	16	95.6%	93.8%	15	21	20
		MLE	41	-28	89	114	96	123	11	16	90.3%	91.0%	-33	85	59
0.3	0.1	Proposed	-4	4	89	120	92	117	8	14	95.1%	95.6%	-12	32	19
		MLE	-4	5	85	121	92	118	9	14	93.8%	95.3%	-155	11	72
0.3	0.3	Proposed	-5	6	91	120	91	121	8	15	95.6%	94.7%	-17	6	9
		MLE	-4	5	90	120	92	121	9	15	94.4%	94.6%	-48	-17	96
0.3	0.5	Proposed	-6	7	95	119	95	118	9	14	94.6%	95.4%	-13	5	7
		MLE	1	1	94	118	97	118	9	14	93.4%	95.2%	-43	36	98
0.5	0.1	Proposed	-6	3	97	119	98	120	10	14	95.2%	95.2%	-8	17	15
		MLE	13	-10	91	118	118	134	14	18	88.2%	91.0%	-150	7	75
0.5	0.3	Proposed	-6	2	99	118	104	119	11	14	93.6%	94.7%	-16	4	6
		MLE	3	-6	98	117	110	122	12	15	92.1%	93.1%	-56	50	97
0.7	0.1	Proposed	-8	4	105	118	109	123	12	15	93.0%	93.5%	-11	5	13
		MLE	97	-89	102	107	126	123	25	23	78.6%	79.5%	90	6	63

Table 3: Simulation results for estimating (β_0, β_1) in a two-level clustering setting with a misspecified joint model but a correct marginal model. Bias, SSE, SEE, MSE represent the same quantities as in Table 1. Two SEE's are presented, one is model-based while the other is robust. 95% confidence interval coverage rates are presented, derived by model based s.e. and robust s.e., respectively.

Two-level clustering, mis-specified conditional model.														
Method	Bias × 10 ³		SEE × 10 ³		Robust SEE × 10 ³		SSE × 10 ³		MSE × 10 ³		95% C.I. coverage rate		Robust 95% C.I. coverage rate	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
Proposed	9	16	133	150	153	151	155	148	24	22	91.1%	95.5%	94.8%	95.9%
MLE	-308	201	144	164	-100	-100	159	153	121	64	42.0%	81.4%	-	-

Table 4: Analysis of Madras longitudinal schizophrenia study.

Coefficients	Exchangeable		AR (1)	
	$\exp(\beta)$	95% C.I.	$\exp(\beta)$	95% C.I.
Likelihood				
Intercept	2.29	(1.44, 3.66)	2.27	(1.43, 3.61)
Time	0.70	(0.66, 0.75)	0.70	(0.66, 0.75)
Age ≤ 20	1.50	(0.83, 2.72)	1.31	(0.73, 2.34)
Female	0.43	(0.24, 0.79)	0.45	(0.26, 0.79)
ρ		0.94		0.95
Proposed Method				
Intercept	2.41	(1.54, 3.78)	2.49	(1.57, 3.93)
Time	0.71	(0.67, 0.75)	0.71	(0.67, 0.76)
Age ≤ 20	1.60	(0.88, 2.90)	1.47	(0.81, 2.66)
Female	0.53	(0.30, 0.95)	0.54	(0.30, 0.96)
ρ		0.92		0.96

Table 5: Analysis of British Social Attitudes Panel Survey: years 1983-1986.

Coefficients	Method 1		Method 2		Method 3	
	$\exp(\beta)$	95% C.I.	$\exp(\beta)$	95% C.I.	$\exp(\beta)$	95% C.I.
Intercept	0.61	(0.23, 1.63)	0.62	(0.24, 1.61)	0.74	(0.22, 2.43)
Year 1984	0.66	(0.49, 0.88)	0.65	(0.48, 0.88)	0.65	(0.47, 0.91)
Year 1985	1.06	(0.80, 1.41)	1.05	(0.78, 1.40)	1.04	(0.74, 1.46)
Year 1986	1.21	(0.91, 1.61)	1.20	(0.90, 1.61)	1.20	(0.88, 1.63)
Class: upper working	0.76	(0.51, 1.14)	0.75	(0.50, 1.13)	0.72	(0.41, 1.24)
Class: lower working	0.80	(0.54, 1.19)	0.78	(0.52, 1.16)	0.66	(0.43, 1.02)
Gender	0.72	(0.48, 1.07)	0.72	(0.49, 1.07)	0.71	(0.45, 1.11)
Religion: catholic	0.67	(0.25, 1.80)	0.67	(0.26, 1.76)	0.76	(0.30, 1.91)
Religion: other	0.52	(0.24, 1.11)	0.52	(0.25, 1.08)	0.45	(0.23, 0.87)
Religion: none	2.00	(1.21, 3.30)	2.02	(1.23, 3.29)	2.12	(1.13, 3.97)
% protestant	2.17	(0.86, 5.52)	2.19	(0.88, 5.48)	1.94	(0.70, 5.41)